Abstract Title Page
Not included in page count.

Title: The Impact of Indiana's System of Diagnostic Assessments on Mathematics Achievement

Author(s): Spyros Konstantopoulos, Shazia Miller, Arie van der Ploeg, Cheng-Hsien Li, Anne Traynor

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Diagnostic assessments are considered by many educational leaders, in states and districts throughout the United States, as an effective lever to increase student achievement. These types of assessments aim to improve ongoing classroom instruction and provide feedback on student performance in order to promote learning for all students. The underlying principle is that via up to date diagnostic information teachers will offer ongoing constructive feedback and individualized instruction that meets the students' needs and improves student learning.

Typically diagnostic assessments are implemented in entire schools and, thus, are considered whole school interventions. Currently, there is a need to evaluate rigorously whole school interventions at a large scale and examine the consistency of the intervention at the state level. The state of Indiana was the first to implement statewide technology-supported interim diagnostic assessments to be taken by all K–8 students multiple times each school year. Indiana expects teachers to use the constantly updated diagnostic information to improve ongoing instruction and increase student achievement. In particular, in 2008 the Indiana Department of Education (IDOE) began the roll-out of its system of diagnostic assessments. The theory undergirding this decision was based on the premise that education is about decreasing gaps between a student's current and intended knowledge. Assessment may be viewed as the measurement of these gaps, providing critical feedback and more exact documentation of the gaps between current and desired status, with repeated measurements documenting changing gaps.

In this study we provide evidence of the effectiveness of diagnostic assessments on mathematics achievement in Indiana schools. We designed a randomized experimental to test the hypothesis that teachers who have access to objective data to monitor student progress to guide their choices about day-to-day instruction will produce students who perform better on state assessments. As a result, we used data from numerous schools in Indiana that implemented the intervention and evaluated whether or not diagnostic assessments lead to improved instructional practices and student outcomes. The results of this study are of practical significance to policymakers, given their current level of interest in assessment levers to accelerate school improvement.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The purpose of the study was to design a rigorous experimental study and collect high quality data to determine the effectiveness of the intervention on student achievement. In particular, we examined whether diagnostic assessments implemented by schools in Indiana produced rigorous estimates of their effects on student performance on the state's annual Indiana Statewide Testing for Educational Progress–Plus (ISTEP+) measures. Because the data were produced from a well designed randomized experiment they have high internal validity and justify causal inferences about the intervention effects (Shadish, Cook, & Campbell, 2002). In addition, because our sample included numerous schools (58 schools) from different school districts in Indiana our results should have higher external validity than those from convenient

and localized samples. To account for the nesting of students within schools we employed two-level models to examine the effectiveness of the intervention.

Setting:
Description of the research location.

The experiment took place in Indiana in 2009-2010 and included K-8 schools that had volunteered to be part of the intervention the Spring of 2009.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features or characteristics.

The RCT is a two-level cluster randomized design. Students were nested within schools, and schools were nested within treatment and control groups. Randomization occurred at the school level—that is, schools were randomly assigned to treatment and control conditions. Since the intervention was designed as a whole-school implementation, the conceptual match between design and practice was good.

Participating schools volunteered to participate in late spring of 2009. From that list of schools, our initial objective was to assign 25 schools to a treatment and 25 to a control group. However, to facilitate participation we decided to use an unbalanced design with a larger number of schools in the treatment group. Specifically our sample included 58 schools, 35 of which were randomly assigned to the treatment condition, while the remaining 23 schools were randomly assigned to the control condition. Of the 35 treatment schools, 31 participated in the experiment and of the 23 control schools, 18 participated in the experiment for the whole year (49 participating schools altogether). Overall, some 20,000 students participated in the study in 2009-2010. The control schools received the treatment with a one-year delay, and the treatment schools continued to receive the treatment in the second year of the study. This design allowed for multiple comparisons between schools in treatment and control groups across years.

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.
For Track 2, this may include the development and validation of a measurement instrument.

In 2006, the Indiana Legislature charged the Indiana Board of Education (ISBE) and the IDOE to develop a long-term plan for a new assessment system that would be less expensive and less time consuming to measure individual student growth from year to year, and provide diagnostic information to teachers to use to improve ongoing instruction. This new system would be fully online. Among the proposals considered was one for new technology-enabled classroom-based diagnostic assessments for all K-8 students that would provide immediate feedback to the teacher and the student.

The plan required that the assessments be voluntary. In schools that chose to use them, IDOE would cover costs. The plan also tasked IDOE to ensure alignment of test content to Indiana standards and grade-level expectations. IDOE identified two commercial products through a standard public agency request for information, request for proposals, and negotiated bidding process. The first program was Wireless Generation's *mCLASS:Reading 3D* and *mCLASS:Math* as the K–2 solution and the second program was CTB/McGraw-Hill's *Acuity* product for Grades 3–8. From IDOE's perspective, this is a single intervention, a system of periodic

diagnostic assessments that is *consistent*, because students throughout Indiana take the same assessments; *periodic*, because students are tested at the same three time points during the school year statewide; and *diagnostic*, because the assessments identify and report individual learning needs.

Indiana began the roll-out of the assessment program in summer 2008 by training teachers from more than 500 schools teaching some 220,000 K–8 students. These teachers and students used the diagnostic assessments during the 2008–09 school year. Additional schools volunteered to participate in the assessments each of the next several years. IDOE staff expects that essentially all elementary schools and students statewide will be active participants by 2013–14.

With the mCLASS, the screening and diagnostic probes are conducted face-to-face, with student and teacher working together. The student performs language tasks while the teacher records characteristics of the work on a personal digital assistant (PDA). Teachers are guided through the assessments by the PDA and, through the PDA interface, they can immediately view results and compare them to prior performance. Detailed individual and group reports as well as ad hoc queries are available to the classroom teacher and other authorized personnel. In addition, at any point, teachers are able to monitor individual student progress in the classroom using short one-on-one, one-minute probes and then see those results linked to previous results graphically on the PDA screen.

Acuity provides online assessments in reading, mathematics, science, and social studies for grades 3-8. These assessments are 30- to 35-item multiple-choice online tests that can be completed within a class period, usually in group settings. They are closely aligned to Indiana standards and designed to be predictive of ISTEP^{+} results. *Acuity* also permits teachers to construct practice or progress monitoring assessments from extensive banks of aligned items for more frequent progress monitoring. Instructional resources—packaged student exercises to practice weak skills or explore others— are also available and may be assigned to specific students directly from *Acuity*'s diagnostic displays. Teacher access to most reports and queries is immediate.

Research Design:

Description of research design

The study was a large-scale randomized experiment where Indiana schools that volunteered to participate were randomly assigned to a treatment and a control condition. The schools in the treatment condition received mCLASS and/or Acuity, while the control schools did not receive any treatment and should not have had any assessment program in place or have received a similar treatment the previous year. Because of random assignment of schools to conditions the results are likely to be causal.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.
For Track 2, this may include the use of existing datasets.

We used data from the first year of the experiment for grades K through 8. We conducted analysis on the initial number of schools that were randomly assigned to conditions (Intention to Treat or ITT) and on the participating or TOT (Treatment on Treated) schools. The outcome was mathematics scores of ISTEP+.

To capture the dependency in the data (i.e., students nested within schools) we used two-level models with students at the first level and schools at the second level (Raudebush & Bryk, 2002). The HLM software was used to analyze the data. We conducted several analysis using all the data (grades K-8), using K-2 data and 3-8 data separately, and finally we conducted within grade analyses as well.

In each case we regressed mathematics scores on the treatment variable that was coded as a binary indicator, and other student and school covariates. The regression model for student i in school j is

$$y_{ij} = \beta_{00} + \beta_{10} Treatment_j + \mathbf{X}_{ij}\mathbf{B}_{20} + \mathbf{Z}_j\mathbf{B}_{30} + \mathbf{G}_{ij}\mathbf{B}_{40} + \upsilon_j + \varepsilon_{ij}$$

where y is mathematics scores, β_{00} is the constant term, β_{10} is the estimate of the treatment effect, *Treatment* is a binary indicator for the treatment, **X** is the design matrix of the student level predictors such as race, gender, SES, etc, $\mathbf{B}_{20}$ is a column vector of regression estimates of student predictors, **Z** is a row vector of school level predictors such as percent female, minority, disadvantaged, prior school achievement, etc, $\mathbf{B}_{30}$ is a column vector of regression estimates of school predictors, **G** represents grade fixed effects, $\mathbf{B}_{40}$ is a column vector of grade fixed effects estimates, υ is a school level residual, and ε is a student level residual. The variance of υ captures the nesting of students within schools. We replicated the analysis described above for grades K to 2 and grades 3 to 8 separately to determine Mclass and Acuity effects respectively. We also conducted within grade analyses for grades K through 6 that had sufficient data to estimate the treatment effect. In this analysis the grade dummies were omitted. For grades 4, 5, and 6 we will also run models that included prior student achievement as a covariate. This analysis was not possible to conduct in other grades given that prior scores were not available (or data were sparse).

Findings / Results:

Description of the main findings with specific details.

The preliminary analysis involved tests that checked whether random assignment of schools to conditions was successful for several observed variables at the school level. The random assignment indicated intention to treat. We used t-tests of independent samples to determine significant differences between the two conditions for several school-level observed variables such as proportion of female, minority, disadvantaged, special education, limited English proficiency students as well as prior school achievement. The results indicated that for these variables random assignment was successful. However, analyses that used data of participating schools were somewhat different and indicated some differences between treatment and control conditions that can result in selection bias.

For the primary analysis mathematics scores were standardized, and as a result the regression estimates are mean differences in standard deviation units between treatment and control groups. Positive estimates indicate a positive treatment effect. The results from the K-8 analysis suggested that overall there was a positive treatment effect. The magnitude of the effect was on average slightly larger than one-twentieth of a standard deviation in mathematics and favored the treatment group. The treatment estimates across different specifications however were not significantly different from zero at the .05 level. That is, although the treatment was positive we did not detect any systematic patterns of its effectiveness on mathematics ISTEP+

scores. Sensitivity analysis that omitted grades 7 and 8 (that had sparse data) produced similar findings.

The K-2 analysis produced results that were overall similar to those discussed above. The treatment effect however was close to zero and not significant in mathematics across different specifications. In contrast, the results produced from the 3-8 analysis were somewhat different. Specifically, the treatment effect in mathematics was consistently larger than one-tenth, and sometimes as large as one-sixth, of a standard deviation, and in some specifications it was significantly different from zero at the .05 level.

The within grade analysis yielded some interesting findings. The treatment effect was positive and consistently significant in fifth grade mathematics across all specifications. The magnitude of the treatment effect ranged from one-fifth to one-fourth of standard deviation and was not trivial. In the sixth grade the treatment effect was positive and significant in mathematics in some specifications and the magnitude of the estimate was as large as one-third of a standard deviation.

The results produced using data from the 49 participating schools (TOT estimates) were somewhat different. In particular, the treatment effect in the K-8 analysis was typically significant in mathematics for most specifications. Similarly, the treatment effect in the 3-8 analysis was positive and typically significant in mathematics for most specifications. The discrepancy in the results may indicate some selection bias since without bias the results from the ITT and the TOT analyses should have been similar.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

The findings overall suggested that across all grades (K to 8) the treatment effect was positive, but it was not as large and not significantly different from zero. As a result, it seems that the type of diagnostic assessment that Indiana implemented in the sample of schools we evaluated did not improve mathematics achievement that much. The same conclusion holds for the results using grades K to 2 and 3 to 8. However, the within grade analysis revealed that in fifth grade mathematics the treatment effect was significant and not trivial. The treatment effect was consistently as large as one-fifth of a standard deviation and indicated an important annual gain in mathematics achievement (see Hill, Bloom, Black, & Lipsey, 2008). All other estimates were not consistently significant and depended on the specification. Hence, it is unclear that the intervention had any systematic effects on student achievement except for fifth grade mathematics. The estimates produced from the TOT analysis indicated significant treatment effects. However, it is likely that the TOT estimates are biased due to selection and therefore they are not as internally valid as those produced by the ITT analysis.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models.* Newbury Park, CA: Sage.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.